(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115862682 B
(45)授权公告日 2023.06.20

(21)申请号 202310000609.8

(22)申请日 2023.01.03

(65)同一申请的已公布的文献号
申请公布号 CN 115862682 A

(43)申请公布日 2023.03.28

(73)专利权人 杭州觅睿科技股份有限公司
地址 310052 浙江省杭州市滨江区西兴街道楚天路91号1栋4楼;2栋2,3,4楼

(72)发明人 顾海军 赵刚强 金伟 应红力

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 王华

(51)Int.Cl.
G10L 25/57(2013.01)
G10L 25/30(2013.01)
G10L 25/03(2013.01)
G10L 25/63(2013.01)
G06V 10/764(2022.01)
G06V 10/774(2022.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)

(56)对比文件
CN 114582355 A,2022.06.03
CN 115019824 A,2022.09.06
肖易明等.引入注意力机制的视频声源定位.信号处理.2019,全文.
A.Vahedian et al..Identification of sound source in machine vision sequences using audio information. Proceedings EC-VIP-MC 2003. 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications (IEEE Cat. No.03EX667).全文.

审查员 王立华

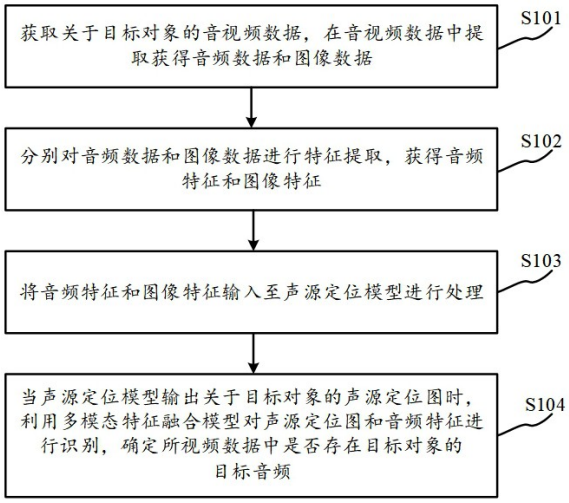
权利要求书2页 说明书13页 附图3页

(54)发明名称

声音检测方法及相关设备

(57)摘要

本申请公开了一种声音检测方法、装置、电子设备及计算机可读存储介质,方法包括:获取关于目标对象的音视频数据,在所述音视频数据中提取获得音频数据和图像数据;分别对所述音频数据和所述图像数据进行特征提取,获得音频特征和图像特征;将所述音频特征和所述图像特征输入至声源定位模型进行处理;当所述声源定位模型输出关于所述目标对象的声源定位图时,利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别,确定所述音视频数据中是否存在所述目标对象的目标音频。应用本申请提供的技术方案,可以有效减少漏检、误检问题,提高声音检测结果的准确性。



CN 115862682 B

1.一种声音检测方法,其特征在于,包括:

获取关于目标对象的音视频数据,在所述音视频数据中提取获得音频数据和图像数据;

分别对所述音频数据和所述图像数据进行特征提取,获得音频特征和图像特征;

将所述音频特征和所述图像特征输入至声源定位模型进行处理;

当所述声源定位模型输出关于所述目标对象的声源定位图时,利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别,确定所述音视频数据中是否存在所述目标对象的目标音频;

其中,所述利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别,包括:

判断是否接收到定制信息,所述定制信息为关于所述目标对象的目标音视频样本;

若是,则利用所述目标音视频样本对所述多模态特征融合模型进行模型优化,获得优化后的多模态特征融合模型;

利用所述优化后的多模态特征融合模型对所述声源定位图和所述音频特征进行识别。

2.根据权利要求1所述的声音检测方法,其特征在于,所述分别对所述音频数据和所述图像数据进行特征提取,获得音频特征和图像特征,包括:

计算所述音频数据的频谱系数,利用音频特征提取模型对所述频谱系数进行特征提取,获得所述音频特征;

利用图像特征提取模型对所述图像数据进行特征提取,获得所述图像特征。

3.根据权利要求1所述的声音检测方法,其特征在于,所述声源定位模型的构建过程包括:

获取音视频样本,并在所述音视频样本中提取得到正音频样本、负音频样本、正图像样本、负图像样本;

对各所述正音频样本进行识别,获得音量值;

将所述音量值不低于预设阈值的正音频样本与所述正图像样本组合为强正样本;

将所述负音频样本和所述负图像样本组合为负样本;

利用所述强正样本和所述负样本对初始声源定位模型进行训练,获得所述声源定位模型。

4.根据权利要求3所述的声音检测方法,其特征在于,所述多模态特征融合模型的构建过程包括:

利用所述声源定位模型对各所述强正样本和各所述负样本进行处理,获得各处理结果,并确定各所述处理结果对应的先验参数;所述处理结果包括输出关于所述目标对象的第一声源定位图、输出关于其他对象的第二声源定位图、无输出;

当所述强正样本的处理结果为输出所述第一声源定位图像时,将所述第一声源定位图像和所述强正样本中的正音频样本组合为第一正样本;

当所述强正样本的处理结果为输出所述第二声源定位图像或所述无输出时,获取所述强正样本中正图像样本的目标对象标定结果,将所述目标对象标定结果和所述强正样本中的正音频样本组合为第二正样本;

当所述负样本的处理结果为输出所述第一声源定位图像时,将所述第一声源定位图和

所述负样本中的负音频样本组合为第一负样本；

当所述负样本的处理结果为输出所述第二声源定位图时，将所述第二声源定位图和所述负样本中的负音频样本组合为第二负样本；

当所述负样本的处理结果为无输出时，获取所述负样本中负图像样本中的其他对象标定结果，将所述其他对象标定结果和所述负样本中的负音频样本组合为第三负样本；

将所述第一正样本、第二正样本组合为正样本集合，将所述第一负样本、第二负样本、第三负样本组合为负样本集合；

根据所述正样本集合、所述负样本集合、各所述先验参数进行模型训练，获得所述多模态特征融合模型。

5.根据权利要求4所述的声音检测方法，其特征在于，还包括：

将所述音量值低于所述预设阈值的正音频样本与所述正图像样本数据组合为弱正样本；

利用所述多模态特征融合模型和所述弱正样本训练获得学生模型；

利用所述学生模型对所述多模态特征融合模型进行参数更新，获得更新后的多模态特征融合模型。

6.根据权利要求1所述的声音检测方法，其特征在于，还包括：

当所述声源定位模型未输出关于所述目标对象的声源定位图时，确定所述音视频数据中不存在所述目标对象的目标音频。

7.一种声音检测装置，其特征在于，包括：

获取模块，用于获取关于目标对象的音视频数据，在所述音视频数据中提取获得音频数据和图像数据；

提取模块，用于分别对所述音频数据和所述图像数据进行特征提取，获得音频特征和图像特征；

输入模块，用于将所述音频特征和所述图像特征输入至声源定位模型进行处理；

识别模块，用于当所述声源定位模型输出关于所述目标对象的声源定位图时，利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别，确定所述音视频数据中是否存在所述目标对象的目标音频；

其中，所述识别模块具体用于判断是否接收到定制信息，所述定制信息为关于所述目标对象的目标音视频样本；若是，则利用所述目标音视频样本对所述多模态特征融合模型进行模型优化，获得优化后的多模态特征融合模型；利用所述优化后的多模态特征融合模型对所述声源定位图和所述音频特征进行识别。

8.一种电子设备，其特征在于，包括：

存储器，用于存储计算机程序；

处理器，用于执行所述计算机程序时实现如权利要求1至6任一项所述的声音检测方法的步骤。

9.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求1至6任一项所述的声音检测方法的步骤。

声音检测方法及相关设备

技术领域

[0001] 本申请涉及多媒体领域，特别涉及一种声音检测方法，还涉及一种声音检测装置、电子设备以及计算机可读存储介质。

背景技术

[0002] 随着生活节奏的加快，对处于婚生阶段的家庭来说，很难做到每时每刻陪伴在婴儿身边，这就会出现无法及时给予婴儿照顾的问题，而婴儿对外界的需求，往往通过哭声来表达，因此，通过智能设备对婴儿进行哭声检测，并能及时反馈父母就显得尤为重要。

[0003] 目前，常见的婴儿哭声检测方法主要是通过提取音频的频谱特征进行判断。但是，这种方法对提取的音频特征要求较高，单一使用音频特征，对于一些易混淆的声音(如猫叫声、木门开门声、鸟叫声、婴儿笑声、小孩尖叫声和交谈声等)容易出现误报现象，对一些声音较小的婴儿哭声，又容易出现漏报。

[0004] 因此，如何有效减少漏检、误检问题，提高声音检测结果的准确性是本领域技术人员亟待解决的问题。

发明内容

[0005] 本申请的目的是提供一种声音检测方法，该声音检测方法可以有效减少漏检、误检问题，提高声音检测结果的准确性；本申请的另一目的是提供一种声音检测装置、电子设备及计算机可读存储介质，均具有上述有益效果。

[0006] 第一方面，本申请提供了一种声音检测方法，包括：

[0007] 获取关于目标对象的音视频数据，在所述音视频数据中提取获得音频数据和图像数据；

[0008] 分别对所述音频数据和所述图像数据进行特征提取，获得音频特征和图像特征；

[0009] 将所述音频特征和所述图像特征输入至声源定位模型进行处理；

[0010] 当所述声源定位模型输出关于所述目标对象的声源定位图时，利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别，确定所述音视频数据中是否存在所述目标对象的目标音频。

[0011] 可选地，所述分别对所述音频数据和所述图像数据进行特征提取，获得音频特征和图像特征，包括：

[0012] 计算所述音频数据的频谱系数，利用音频特征提取模型对所述频谱系数进行特征提取，获得所述音频特征；

[0013] 利用图像特征提取模型对所述图像数据进行特征提取，获得所述图像特征。

[0014] 可选地，所述声源定位模型的构建过程包括：

[0015] 获取音视频样本，并在所述音视频样本中提取得到正音频样本、负音频样本、正图像样本、负图像样本；

[0016] 对各所述正音频样本进行识别，获得音量值；

[0017] 将所述音量值不低于预设阈值的正音频样本与所述正图像样本组合为强正样本；

[0018] 将所述负音频样本和所述负图像样本组合为负样本；

[0019] 利用所述强正样本和所述负样本对初始声源定位模型进行训练，获得所述声源定位模型。

[0020] 可选地，所述多模态特征融合模型的构建过程包括：

[0021] 利用所述声源定位模型对各所述强正样本和各所述负样本进行处理，获得各处理结果，并确定各所述处理结果对应的先验参数；所述处理结果包括输出关于所述目标对象的第一声源定位图、输出关于其他对象的第二声源定位图、无输出；

[0022] 当所述强正样本的处理结果为输出所述第一声源定位图像时，将所述第一声源定位图像和所述强正样本中的正音频样本组合为第一正样本；

[0023] 当所述强正样本的处理结果为输出所述第二声源定位图像或所述无输出时，获取所述强正样本中正图像样本的目标对象标定结果，将所述目标对象标定结果和所述强正样本中的正音频样本组合为第二正样本；

[0024] 当所述负样本的处理结果为输出所述第一声源定位图像时，将所述第一声源定位图和所述负样本中的负音频样本组合为第一负样本；

[0025] 当所述负样本的处理结果为输出所述第二声源定位图时，将所述第二声源定位图和所述负样本中的负音频样本组合为第二负样本；

[0026] 当所述负样本的处理结果为无输出时，获取所述负样本中负图像样本中的其他对象标定结果，将所述其他对象标定结果和所述负样本中的负音频样本组合为第三负样本；

[0027] 将所述第一正样本、第二正样本组合为正样本集合，将所述第一负样本、第二负样本、第三负样本组合为负样本集合；

[0028] 根据所述正样本集合、所述负样本集合、各所述先验参数进行模型训练，获得所述多模态特征融合模型。

[0029] 可选地，所述声音检测方法还包括：

[0030] 将所述音量值低于所述预设阈值的正音频样本与所述正图像样本数据组合为弱正样本；

[0031] 利用所述多模态特征融合模型和所述弱正样本训练获得学生模型；

[0032] 利用所述学生模型对所述多模态特征融合模型进行参数更新，获得更新后的多模态特征融合模型。

[0033] 可选地，所述利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别，包括：

[0034] 判断是否接收到定制信息，所述定制信息为关于所述目标对象的目标音视频样本；

[0035] 若是，则利用所述目标音视频样本对所述多模态特征融合模型进行模型优化，获得优化后的多模态特征融合模型；

[0036] 利用所述优化后的多模态特征融合模型对所述声源定位图和所述音频特征进行识别。

[0037] 可选地，所述声音检测方法还包括：

[0038] 当所述声源定位模型未输出关于所述目标对象的声源定位图时，确定所述音视频

数据中不存在所述目标对象的目标音频。

[0039] 第二方面，本申请还公开了一种声音检测装置，包括：

[0040] 获取模块，用于获取关于目标对象的音视频数据，在所述音视频数据中提取获得音频数据和图像数据；

[0041] 提取模块，用于分别对所述音频数据和所述图像数据进行特征提取，获得音频特征和图像特征；

[0042] 输入模块，用于将所述音频特征和所述图像特征输入至声源定位模型进行处理；

[0043] 识别模块，用于当所述声源定位模型输出关于所述目标对象的声源定位图时，利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别，确定所述音视频数据中是否存在所述目标对象的目标音频。

[0044] 第三方面，本申请还公开了一种电子设备，包括：

[0045] 存储器，用于存储计算机程序；

[0046] 处理器，用于执行所述计算机程序时实现如上所述的任一种声音检测方法的步骤。

[0047] 第四方面，本申请还公开了一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如上所述的任一种声音检测方法的步骤。

[0048] 本申请提供了一种声音检测方法，包括获取关于目标对象的音视频数据，在所述音视频数据中提取获得音频数据和图像数据；分别对所述音频数据和所述图像数据进行特征提取，获得音频特征和图像特征；将所述音频特征和所述图像特征输入至声源定位模型进行处理；当所述声源定位模型输出关于所述目标对象的声源定位图时，利用多模态特征融合模型对所述声源定位图和所述音频特征进行识别，确定所述音视频数据中是否存在所述目标对象的目标音频。

[0049] 应用本申请所提供的技术方案，首先获取音视频数据，并从中分别提取音频数据和图像数据，然后利用声源定位模型对音频数据的音频特征和图像数据的图像特征进行处理获取关于目标对象的声源定位图，最后利用多模态特征融合模型对声源定位图和音频数据的音频特进行处理，以确定音视频数据中是否存在关于目标对象的目标声音，从而实现声音检测，显然，该种实现方式实现了多模态特征的声音检测，相较于单一模态特征的声音检测，可以有效减少漏检、误检问题，从而提高声音检测结果的准确性。

[0050] 本申请所提供的声音检测装置、电子设备以及计算机可读存储介质，同样具有上述技术效果，本申请在此不再赘述。

附图说明

[0051] 为了更清楚地说明现有技术和本申请实施例中的技术方案，下面将对现有技术和本申请实施例描述中需要使用的附图作简要的介绍。当然，下面有关本申请实施例的附图描述的仅仅是本申请中的一部分实施例，对于本领域普通技术人员来说，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图，所获得的其他附图也属于本申请的保护范围。

[0052] 图1为本申请所提供的一种声音检测方法的流程示意图；

[0053] 图2为本申请所提供的一种婴儿哭声检测方法的流程示意图；

[0054] 图3为本申请所提供的一种声音检测装置的结构示意图；

[0055] 图4为本申请所提供的一种电子设备的结构示意图。

具体实施方式

[0056] 本申请的核心是提供一种声音检测方法，该声音检测方法可以有效减少漏检、误检问题，提高声音检测结果的准确性；本申请的另一核心是提供一种声音检测装置、电子设备及计算机可读存储介质，均具有上述有益效果。

[0057] 为了对本申请实施例中的技术方案进行更加清楚、完整地描述，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行介绍。显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0058] 本申请实施例提供了一种声音检测方法。

[0059] 请参考图1，图1为本申请所提供的一种声音检测方法的流程示意图，该声音检测方法可以包括如下S101至S104。

[0060] S101：获取关于目标对象的音视频数据，在音视频数据中提取获得音频数据和图像数据。

[0061] 本步骤旨在实现音视频数据的获取，以及音视频数据中音频数据和图像数据的提取。其中，音视频数据是关于目标对象的音视频数据，可以通过视频采集设备采集获得，目标对象即为需要进行声音检测的对象，例如，当需要进行婴儿哭声检测时，该婴儿即为目标对象。进一步，在音视频数据中提取获得音频数据和图像数据。

[0062] 其中，图像数据是指从音视频数据中提取得到的图像帧，为减少计算量，提高检测效率，可以按照预先设定的时间间隔进行图像数据的少量提取，例如，可以每间隔3秒提取10张图像数据。音频数据是指从音视频数据中提取得到的音频片段，可以为完整的音频片段，同样的，为减少计算量，可以按照预先设定的时间间隔进行短音频片段的提取，例如，可以每间隔3秒钟提取时长为5秒钟的音频数据。当然，提取数据的时间间隔、图像数据的单次提取数量、视频数据的单次提取时长均不影响本技术方案的实施，由技术人员根据实际需求进行设置即可，本申请对此不做限定。

[0063] S102：分别对音频数据和图像数据进行特征提取，获得音频特征和图像特征。

[0064] 本步骤旨在实现音频特征和图像特征的提取。在基于音视频数据获得音频数据和图像数据之后，可以分别对二者进行特征提取，获得相应的音频特征和图像特征，以便于基于这些特征实现后续声音检测。

[0065] 其中，特征提取可以基于相应的网络模型实现，在一种可能的实现方式中，上述分别对音频数据和图像数据进行特征提取，获得音频特征和图像特征，可以包括：计算音频数据的频谱系数，利用音频特征提取模型对频谱系数进行特征提取，获得音频特征；利用图像特征提取模型对图像数据进行特征提取，获得图像特征。

[0066] 可以理解的是，根据人耳的听觉特性可知，人耳对频率具有选择性，且不能有效地分辨出所有的频率分量，因此针对音频数据，使用类似人耳听觉机理的梅尔倒谱系数计算频谱特征能有效表征原始声音特性。在此基础上，对于音频数据而言，可以先计算音频数据

的频谱系数，然后将其输入至音频特征提取模型进行特征提取，即可得到相应的音频特征；对于图像数据而言，则可以直接将其输入至图像特征提取模型进行特征提取，得到相应的图像特征。

[0067] 其中，音频特征提取模型和图像特征提取模型可以使用具有相同框架的深度卷积神经网络训练得到，在一种可能的实现方式中，深度卷积神经网络可以为resnet-18残差网络。

[0068] S103：将音频特征和图像特征输入至声源定位模型进行处理。

[0069] 本步骤旨在实现基于声源定位模型的特征数据处理。具体而言，在得到音频数据的音频特征和图像数据的图像特征之后，即可将其输入至声源定位模型进行处理。其中，声源定位模型是用于实现声源定位的网络模型，可预存于相应的存储空间，在使用时直接调取即可。具体而言，声源定位模型可以通过对音频特征和图像特征进行识别处理，获得声源定位图，声源定位图则用于显示发出声音的各个位置的位置信息，当然，该声音定位图可能为关于目标对象的声源定位图（如展示有婴儿哭声的位置）；也可能为关于其他对象（除目标对象之外）的声源定位图（如展示有宠物叫声的位置）；此外，当然也存在声源定位模型未输出声源定位图的情况，显然，该种情况对应于当前环境较为安静的场景。

[0070] S104：当声源定位模型输出关于目标对象的声源定位图时，利用多模态特征融合模型对声源定位图和音频特征进行识别，确定音视频数据中是否存在目标对象的目标音频。

[0071] 本步骤旨在实现最终的声音检测，以确定音视频数据中是否存在目标对象的目标音频。具体而言，在声源定位模型输出声源定位图，且该声源定位图为关于目标对象的声源定位图的情况下，可以进一步调取多模态特征融合模型，并将声源定位模型输出的声源定位图和音频特征输入该多模态特征融合模型进行二次识别，多模态特征融合模型的输出即为最终的声音检测结果。

[0072] 可以理解的是，基于多模态特征融合模型，不仅实现了多模态特征的声音检测，也实现了二次声音检测，可以使得最终的声音检测结果得到显著提升。例如，在婴儿哭声检测场景下，基于声源定位模型虽然能够解决大部分类似婴儿哭声的误报，但对于一些婴儿自身发出的声音（如婴儿短促的喊叫、笑声、说话声等），大概率会定位到婴儿身上，无法准确细分婴儿是否哭泣而产生误检；在某些特殊音视频场景中，音频声音比较小或者没有明显声音时（如麦克风有故障时），但视频图像中确实有婴儿哭泣，声源定位模型几乎不产生效果，从而产生漏检。针对此类问题，本申请将声源定位模型和多模态特征融合模型相结合实现声音检测，显然可以大大减少漏检、误检的问题。

[0073] 如上所述，基于声源定位模型的特征数据处理同样存在声源定位模型未输出声源定位图的情况，显然该种情况对应于当前环境较为安静的场景，因此，在本申请的一个实施例中，该声音检测方法还可以包括：当声源定位模型未输出关于目标对象的声源定位图时，确定音视频数据中不存在目标对象的目标音频。

[0074] 可见，本申请实施例所提供的声音检测方法，首先获取音视频数据，并从中分别提取音频数据和图像数据，然后利用声源定位模型对音频数据的音频特征和图像数据的图像特征进行处理获取关于目标对象的声源定位图，最后利用多模态特征融合模型对声源定位图和音频数据的音频特进行处理，以确定音视频数据中是否存在关于目标对象的目标声

音，从而实现声音检测，显然，该种实现方式实现了多模态特征的声音检测，相较于单一模态特征的声音检测，可以有效减少漏检、误检问题，从而提高声音检测结果的准确性。

[0075] 在上述实施例的基础上：

[0076] 在本申请的一个实施例中，声源定位模型的构建过程可以包括：

[0077] 获取音视频样本，并在音视频样本中提取得到正音频样本、负音频样本、正图像样本、负图像样本；

[0078] 对各正音频样本进行识别，获得音量值；

[0079] 将音量值不低于预设阈值的正音频样本与正图像样本组合为强正样本；

[0080] 将负音频样本和负图像样本组合为负样本；

[0081] 利用强正样本和负样本对初始声源定位模型进行训练，获得声源定位模型。

[0082] 本申请实施例提供了一种声源定位模型的构建方法。首先，获取音视频样本，为实现模型训练，此处应当使用较多数量的音视频样本，对于每一个音视频样本，均对其进行正音频样本、负音频样本、正图像样本、负图像样本的提取，其中，正音频样本是指包含有目标对象的目标声音(此处指定声音类型旨在实现对应类型的声音检测)的音频样本，负音频样本是指包含有除目标对象之外的其他对象的声音信息的音频样本，正图像样本是指包含有目标对象、且目标对象正在执行某动作导致发出目标声音的图像样本，负图像样本是指不包含有目标对象或仅包含有目标对象但目标对象并未执行某动作导致发出目标声音的图像样本。进一步，对于每一个正音频样本，均可以对其进行音量识别，得到音量值不低于预设阈值的正音频样本，并将音量值不低于预设阈值的正音频样本与正图像样本组合为强正样本，作为训练声源定位模型的正样本，将负音频样本和负图像样本组合为负样本，作为训练声源定位模型的负样本；其中，预设阈值的取值由技术人员根据实际需求进行设置即可，本申请对此不做限定。最后，利用强正样本和负样本对初始声源定位模型进行训练，即可获得声源定位模型，其中，初始声源定位模型可以为基于神经网络模型搭建的初始模型，通过强正样本和负样本对其进行训练即可得到最终的声源定位模型，也可以为传统技术中常用的声源定位模型，通过强正样本和负样本对其进行参数调优即可得到最终的声源定位模型。

[0083] 在本申请的一个实施例中，多模态特征融合模型的构建过程可以包括：

[0084] 利用声源定位模型对各强正样本和各负样本进行处理，获得各处理结果，并确定各处理结果对应的先验参数；处理结果包括输出关于目标对象的第一声源定位图、输出关于其他对象的第二声源定位图、无输出；

[0085] 当强正样本的处理结果为输出第一声源定位图像时，将第一声源定位图像和强正样本中的正音频样本组合为第一正样本；

[0086] 当强正样本的处理结果为输出第二声源定位图像或无输出时，获取强正样本中正图像样本的目标对象标定结果，将目标对象标定结果和强正样本中的正音频样本组合为第二正样本；

[0087] 当负样本的处理结果为输出第一声源定位图像时，将第一声源定位图和负样本中的负音频样本组合为第一负样本；

[0088] 当负样本的处理结果为输出第二声源定位图时，将第二声源定位图和负样本中的负音频样本组合为第二负样本；

[0089] 当负样本的处理结果为无输出时，获取负样本中负图像样本中的其他对象标定结果，将其他对象标定结果和负样本中的负音频样本组合为第三负样本；

[0090] 将第一正样本、第二正样本组合为正样本集合，将第一负样本、第二负样本、第三负样本组合为负样本集合；

[0091] 根据正样本集合、负样本集合、各先验参数进行模型训练，获得多模态特征融合模型。

[0092] 本申请实施例提供了一种多模态特征融合模型的构建方法，该多模态特征融合模型的训练样本基于声源定位模型对样本数据的处理结果生成。首先，在完成声源定位模型训练之后，可以先利用该声源定位模型分别对每一个强正样本和每一个负样本进行处理，获得每一个样本数据对应的处理结果，该处理结果可能分为如下三种情况：输出关于目标对象的第一声源定位图、输出关于其他对象的第二声源定位图、无输出；在此基础上，根据每一个处理结果确定及对应的先验参数，该先验参数即为用于实现多模态特征融合模型训练的参数信息，并且，对于不同的处理结果，先验参数对应于不同的取值，即先验参数由处理结果所决定。进一步，针对不同的处理结果，划分用于进行多模态特征融合模型训练的样本数据集，包括正样本数据集合负样本数据集：

[0093] 1、对于强正样本而言：

[0094] 若其处理结果为输出第一声源定位图像，则将该第一声源定位图像和该强正样本中的正音频样本组合为第一正样本；

[0095] 若其处理结果为输出第二声源定位图像或者为无输出，则由技术人员对强正样本中的正图像样本进行人工标定，获得关于目标对象的标定结果，并将该目标对象标定结果和该强正样本中的正音频样本组合为第二正样本。

[0096] 最后，所有的第一正样本和所有的第二正样本组合为正样本集合，作为训练多模态特征融合模型的正样本。

[0097] 2、对于负样本而言：

[0098] 若其处理结果为输出第一声源定位图像，则将该第一声源定位图和该负样本中的负音频样本组合为第一负样本；

[0099] 若其处理结果为输出第二声源定位图像，则将该第二声源定位图和该负样本中的负音频样本组合为第二负样本；

[0100] 若其处理结果为无输出，则由技术人员对负样本中的负样本图像进行人工标定，获得关于其他对象的标定结果，并将该其他对象标定结果和该负样本中的负音频样本组合为第三负样本。

[0101] 最后，所有的第一负样本、所有的第二负样本、所有的第三负样本组合为负样本集合，作为训练多模态特征融合模型的负样本。

[0102] 由此，即可利用正样本集合、负样本集合以及各样本集合内每一个样本对应的先验参数进行模型训练，得到最终的多模态特征融合模型。

[0103] 在本申请的一个实施例中，该声音检测方法还可以包括：

[0104] 将音量值低于预设阈值的正音频样本与正图像样本数据组合为弱正样本；

[0105] 利用多模态特征融合模型和弱正样本训练获得学生模型；

[0106] 利用学生模型对多模态特征融合模型进行参数更新，获得更新后的多模态特征融

合模型。

[0107] 为有效提高模型精度，进一步提高声音检测结果的准确性，可以继续对上一实施例训练得到的多模态特征融合模型进行参数更新，获得同时适用于高音量场景和低音量场景的多模态特征融合模型。

[0108] 具体而言，在构建声源定位模型的过程中，对于音量值低于预设阈值的正音频样本，可将其与正图像样本组合为弱正样本，以便于利用该弱正样本对多模态特征融合模型继续进行参数更新。首先，将上述获得的多模态特征融合模型作为教师模型，结合弱正样本训练得到一个学生模型；然后，利用该学生模型通过指数移动平均值的方式对教师模型中的参数进行更新，即可得到更新后的教师模型，也即上述更新后的多模态特征融合模型。显然，由于更新后的多模态特征融合模型是基于包含有音量值低于预设阈值的正音频样本的弱正样本更新得到，而上一实施例获得的多模态特征融合模型是基于包含有音量值不低于预设阈值的正音频样本的强正样本训练得到，因此，该更新后的多模态特征融合模型可同时适用于高音量场景和低音量场景，保证声音检测结果的准确性。

[0109] 在本申请的一个实施例中，上述利用多模态特征融合模型对声源定位图和音频特征进行识别，可以包括：

[0110] 判断是否接收到定制信息，定制信息为关于目标对象的目标音视频样本；

[0111] 若是，则利用目标音视频样本对多模态特征融合模型进行模型优化，获得优化后的多模态特征融合模型；

[0112] 利用优化后的多模态特征融合模型对声源定位图和音频特征进行识别。

[0113] 本申请实施例提供了一种基于定制化多模态特征融合模型进行声音检测的实现方法，可以满足不同用户的定制化需求，且可以提高声音检测结果的准确性。

[0114] 可以理解的是，用于进行多模态特征融合模型训练的样本数据来源各不相同，例如，在婴儿哭声检测场景下，用于进行多模态特征融合模型训练的样本数据必然是来自于不同婴儿的哭声样本，在此基础上，为实现针对某一特定目标对象的声音检测，可以获取包含有关于该特定目标对象的目标音视频样本的定制信息，并基于这些目标音视频样本对多模态特征融合模型进行模型优化，得到优化后的多模态特征融合模型，此时，该优化后的多模态特征融合模型必然是最适用于前述特定目标对象的网络模型，最后，再利用该优化后的多模态特征融合模型继续进行声音检测，显然，其检测结果具备更高的准确性。

[0115] 在上述各实施例的基础上，本申请实施例提供了另一种声音检测方法。本申请实施例所提供的声音检测方法以婴儿哭声检测为例，其实现流程如下：

[0116] 一、模型训练：

[0117] 1、样本数据获取：

[0118] 收集家用场景中出现婴儿哭泣的音视频，进行人工挑选，建立数据库。对收集的音视频数据，可以每间隔3秒做一次截断，若该段音视频中含有婴儿哭声，且同时含有婴儿正在哭泣的图像，则在该段时间内，挑选一张婴儿正在哭泣时的图像，与音频形成配对，组成一对强正样本。若该段声音音量很小或无声，但视频中婴儿确实在哭泣，则挑选一张婴儿正在哭泣时的图像并对面部进行标定，将标定图与该音频形成配对，组成一对弱正样本。同时，还需收集类似婴儿哭声的对比训练样本集(也可用于负样本)，包括猫叫、木门开关、火车行驶、鸟叫等常见场景下的音视频配对数据集，和各类非婴儿哭泣的图像作为负样本。

[0119] 2、特征数据提取：

[0120] 对于各样本数据中的音频样本，计算其梅尔倒谱系数，并使用深度卷积神经网络对其进行特征提取，获取对应的音频特征。对于各样本数据中的图像样本，使用具有相同框架的深度卷积神经网络进行特征提取，获取对应的图像特征。

[0121] 3、声源定位模型训练及其数据处理：

[0122] 使用强正样本数据集和负样本数据集对EZ-VSL算法中提供的初始声源定位模型进行网络参数微调，获得最终的声源定位模型。

[0123] 利用训练获得的声源定位模型对各强正样本数据和负样本数据进行婴儿声源定位，获得相应的定位结果。

[0124] 4、定位结果分类：

[0125] 如若声源定位模型输出有声源定位图，则先对其进行等比例缩放、补边等预处理，然后送入深度卷积神经网络进行分类，判断该声源定位图是否为婴儿的声源定位图。如若声源定位模型无输出，则可以根据实际情况进行下一步判断。

[0126] (1)针对强正样本：

[0127] (a)当输出声源定位图时，若声源定位图的分类置信度score大于等于预设阈值，即判定为是关于婴儿的声源定位图时，则令$\alpha=\text{score}$，不做其他处理；若声源定位图的分类置信度score小于预设阈值，即判断不是关于婴儿的声源定位图时，则令$\alpha=1$，获取视频中关于婴儿的人工标定结果。然后，将上述两种情况的音频样本和定位结果(声源定位图或人工标定结果)组合为正样本。

[0128] (b)当未输出声源定位图时，则令$\alpha=1$，获取视频中关于婴儿的人工标定结果，将该人工标定结果和相应的音频样本组合为正样本。

[0129] 上述获得的所有正样本组合为用于后续训练多模态特征融合模型的正样本数据。

[0130] (2)针对负样本：

[0131] (a)当输出声源定位图时，若声源定位图的分类置信度score大于等于预设阈值，即判定为是关于婴儿的声源定位图时，则令$\alpha=0$，不做其他处理；若声源定位图的分类置信度score小于预设阈值，即判断不是关于婴儿的声源定位图时，则令$\alpha=1$。然后，将上述两种情况的音频样本和声源定位图组合为负样本。

[0132] (b)当未输出声源定位图时，则令$\alpha=1$，获取视频中关于非婴儿的人工标定结果，将人工标定结果和相应的音频样本组合为负样本。

[0133] 上述获得的所有负样本组合为用于后续训练多模态特征融合模型的负样本数据。

[0134] 其中，α为后续用于进行多模态特征融合模型训练的损失函数提供的先验参数。

[0135] 5、多模态特征融合模型训练及其数据处理：

[0136] 对于每一正样本、每一负样本，分别使用深度卷积神经网络进行特征提取，并使用如下公式作为loss函数进行多模态特征融合模型训练；

[0137] $$loss=\frac{\alpha}{2}loss_a+(1-\frac{\alpha}{2})loss_v$$；

[0138] 其中，$loss_a$为音频损失函数，$loss_v$为图像损失函数，二者均使用二元交叉熵。

[0139] 通过对样本数据集进行数据清理，获取干净数据集进行训练，动态调整音频损失

和图像损失对总损失的占比，利用随机梯度下降算法对总损失loss进行优化求解，更新融合的网络参数，即可得到最优的预测模型，即多模态特征融合模型。

[0140] 进一步，为进一步提高模型精度，针对弱声音或无声音场景，可以将上述预测模型作为教师模型指导训练一个具有相同模型架构的学生模型，公式如下：

[0141] $loss_{student} = loss_{\det} + \lambda loss_{consist}$；

[0142] $loss_{\det} = loss(X_a, X_v, Y)$；

[0143] $loss_{consist} = loss(X_a, X_v, \hat{C}_t)$；

[0144] 其中，X_a 表示输入的音频特征，X_v 表示输入的图像特征，Y 表示输入的婴儿哭声标签，$\hat{C}_t$ 表示教师模型对输入音视频特征的预测，λ 为训练超参数，实际使用时 $\lambda = 1$。

[0145] 由此，学生模型则可以通过指数移动平均值的方式更新教师模型中的参数，达到模型之间的互相学习，使用公式如下：

[0146] $w_t \leftarrow \beta w_t + (1-\beta) w_s$；

[0147] 其中，w_t 为教师模型参数，w_s 为学生模型参数，$\beta = 0.996$。

[0148] 至此，得到更新后的教师模型，也即更新后的多模态特征融合模型。

[0149] 二、基于训练模型的婴儿哭声检测：

[0150] 请参考图2，图2为本申请所提供的一种婴儿哭声检测方法的流程示意图，其实现流程如下：

[0151] 1、获取关于目标婴儿的音视频数据，并分别提取得到音频数据和视频数据；

[0152] 2、分别对音频数据和视频数据进行特征提取，得到音频特征和视频特征；

[0153] 3、将音频特征和视频特征输入至声源定位模型进行处理；

[0154] 4、当声源定位模型未输出声源定位图时，可以确定音视频数据中不存在目标婴儿的哭声；

[0155] 5、当声源定位模型输出声源定位图时，对该声源定位图进行婴儿识别，判断其是否为关于目标婴儿的声源定位图；

[0156] 6、当声源定位图为关于婴儿的声源定位图时，将其与音频特征进行融合，得到融合特征；

[0157] 7、确定当前是否接收到用户的定制信息，若否，则直接调取多模态特征融合模型进行识别处理，获得声音检测结果，从而确定音视频数据中是否存在目标婴儿的哭声；

[0158] 8、若接收到了用户的定制信息，则利用目标婴儿的音视频样本对多模态特征融合模型进行参数微调，实现模型自适应；

[0159] 9、利用自适应后的多模态特征融合模型(即上述优化后的多模态特征融合模型)对融合特征进行识别处理，获得声音检测结果，从而确定音视频数据中是否存在目标婴儿的哭声。

[0160] 可见，本申请实施例所提供的声音检测方法，首先获取音视频数据，并从中分别提取音频数据和图像数据，然后利用声源定位模型对音频数据的音频特征和图像数据的图像

特征进行处理获取关于目标对象的声源定位图，最后利用多模态特征融合模型对声源定位图和音频数据的音频特进行处理，以确定音视频数据中是否存在关于目标对象的目标声音，从而实现声音检测，显然，该种实现方式实现了多模态特征的声音检测，相较于单一模态特征的声音检测，可以有效减少漏检、误检问题，从而提高声音检测结果的准确性。

[0161] 本申请实施例提供了一种声音检测装置。

[0162] 请参考图3，图3为本申请所提供的一种声音检测装置的结构示意图，该声音检测装置可以包括：

[0163] 获取模块1，用于获取关于目标对象的音视频数据，在音视频数据中提取获得音频数据和图像数据；

[0164] 提取模块2，用于分别对音频数据和图像数据进行特征提取，获得音频特征和图像特征；

[0165] 输入模块3，用于将音频特征和图像特征输入至声源定位模型进行处理；

[0166] 识别模块4，用于当声源定位模型输出关于目标对象的声源定位图时，利用多模态特征融合模型对声源定位图和音频特征进行识别，确定音视频数据中是否存在目标对象的目标音频。

[0167] 可见，本申请实施例所提供的声音检测装置，首先获取音视频数据，并从中分别提取音频数据和图像数据，然后利用声源定位模型对音频数据的音频特征和图像数据的图像特征进行处理获取关于目标对象的声源定位图，最后利用多模态特征融合模型对声源定位图和音频数据的音频特进行处理，以确定音视频数据中是否存在关于目标对象的目标声音，从而实现声音检测，显然，该种实现方式实现了多模态特征的声音检测，相较于单一模态特征的声音检测，可以有效减少漏检、误检问题，从而提高声音检测结果的准确性。

[0168] 在本申请的一个实施例中，上述提取模块2可具体用于计算音频数据的频谱系数，利用音频特征提取模型对频谱系数进行特征提取，获得音频特征；利用图像特征提取模型对图像数据进行特征提取，获得图像特征。

[0169] 在本申请的一个实施例中，该声音检测装置还可以包括第一模型构建模块，用于获取音视频样本，并在音视频样本中提取得到正音频样本、负音频样本、正图像样本、负图像样本；对各正音频样本进行识别，获得音量值；将音量值不低于预设阈值的正音频样本与正图像样本组合为强正样本；将负音频样本和负图像样本组合为负样本；利用强正样本和负样本对初始声源定位模型进行训练，获得声源定位模型。

[0170] 在本申请的一个实施例中，该声音检测装置还可以包括第二模型构建模块，用于利用声源定位模型对各强正样本和各负样本进行处理，获得各处理结果，并确定各处理结果对应的先验参数；处理结果包括输出关于目标对象的第一声源定位图、输出关于其他对象的第二声源定位图、无输出；当强正样本的处理结果为输出第一声源定位图像时，将第一声源定位图像和强正样本中的正音频样本组合为第一正样本；当强正样本的处理结果为输出第二声源定位图像或无输出时，获取强正样本中正图像样本的目标对象标定结果，将目标对象标定结果和强正样本中的正音频样本组合为第二正样本；当负样本的处理结果为输出第一声源定位图像时，将第一声源定位图和负样本中的负音频样本组合为第一负样本；当负样本的处理结果为输出第二声源定位图时，将第二声源定位图和负样本中的负音频样本组合为第二负样本；当负样本的处理结果为无输出时，获取负样本中负图像样本中的其

他对象标定结果，将其他对象标定结果和负样本中的负音频样本组合为第三负样本；将第一正样本、第二正样本组合为正样本集合，将第一负样本、第二负样本、第三负样本组合为负样本集合；根据正样本集合、负样本集合、各先验参数进行模型训练，获得多模态特征融合模型。

[0171] 在本申请的一个实施例中，该声音检测装置还可以包括模型更新模块，用于将音量值低于预设阈值的正音频样本与正图像样本数据组合为弱正样本；利用多模态特征融合模型和弱正样本训练获得学生模型；利用学生模型对多模态特征融合模型进行参数更新，获得更新后的多模态特征融合模型。

[0172] 在本申请的一个实施例中，上述识别模块4可具体用于判断是否接收到定制信息，定制信息为关于目标对象的目标音视频样本；若是，则利用目标音视频样本对多模态特征融合模型进行模型优化，获得优化后的多模态特征融合模型；利用优化后的多模态特征融合模型对声源定位图和音频特征进行识别。

[0173] 在本申请的一个实施例中，上述识别模块4还可用于当声源定位模型未输出关于目标对象的声源定位图时，确定音视频数据中不存在目标对象的目标音频。

[0174] 对于本申请实施例提供的装置的介绍请参照上述方法实施例，本申请在此不做赘述。

[0175] 本申请实施例提供了一种电子设备。

[0176] 请参考图4，图4为本申请所提供的一种电子设备的结构示意图，该电子设备可包括：

[0177] 存储器，用于存储计算机程序；

[0178] 处理器，用于执行计算机程序时可实现如上述任意一种声音检测方法的步骤。

[0179] 如图4所示，为电子设备的组成结构示意图，电子设备可以包括：处理器10、存储器11、通信接口12和通信总线13。处理器10、存储器11、通信接口12均通过通信总线13完成相互间的通信。

[0180] 在本申请实施例中，处理器10可以为中央处理器(Central Processing Unit，CPU)、特定应用集成电路、数字信号处理器、现场可编程门阵列或者其他可编程逻辑器件等。

[0181] 处理器10可以调用存储器11中存储的程序，具体的，处理器10可以执行声音检测方法的实施例中的操作。

[0182] 存储器11中用于存放一个或者一个以上程序，程序可以包括程序代码，程序代码包括计算机操作指令，在本申请实施例中，存储器11中至少存储有用于实现以下功能的程序：

[0183] 获取关于目标对象的音视频数据，在音视频数据中提取获得音频数据和图像数据；

[0184] 分别对音频数据和图像数据进行特征提取，获得音频特征和图像特征；

[0185] 将音频特征和图像特征输入至声源定位模型进行处理；

[0186] 当声源定位模型输出关于目标对象的声源定位图时，利用多模态特征融合模型对声源定位图和音频特征进行识别，确定音视频数据中是否存在目标对象的目标音频。

[0187] 在一种可能的实现方式中，存储器11可包括存储程序区和存储数据区，其中，存储

程序区可存储操作系统,以及至少一个功能所需的应用程序等;存储数据区可存储使用过程中所创建的数据。

[0188] 此外,存储器11可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件或其他易失性固态存储器件。

[0189] 通信接口12可以为通信模块的接口,用于与其他设备或者系统连接。

[0190] 当然,需要说明的是,图4所示的结构并不构成对本申请实施例中电子设备的限定,在实际应用中电子设备可以包括比图4所示的更多或更少的部件,或者组合某些部件。

[0191] 本申请实施例提供了一种计算机可读存储介质。

[0192] 本申请实施例所提供的计算机可读存储介质上存储有计算机程序,计算机程序被处理器执行时可实现如上述任意一种声音检测方法的步骤。

[0193] 该计算机可读存储介质可以包括:U盘、移动硬盘、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0194] 对于本申请实施例提供的计算机可读存储介质的介绍请参照上述方法实施例,本申请在此不做赘述。

[0195] 说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0196] 专业人员还可以进一步意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0197] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块,或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM或技术领域内所公知的任意其它形式的存储介质中。

[0198] 以上对本申请所提供的技术方案进行了详细介绍。本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想。应当指出,对于本技术领域的普通技术人员来说,在不脱离本申请原理的前提下,还可以对本申请进行若干改进和修饰,这些改进和修饰也落入本申请的保护范围内。

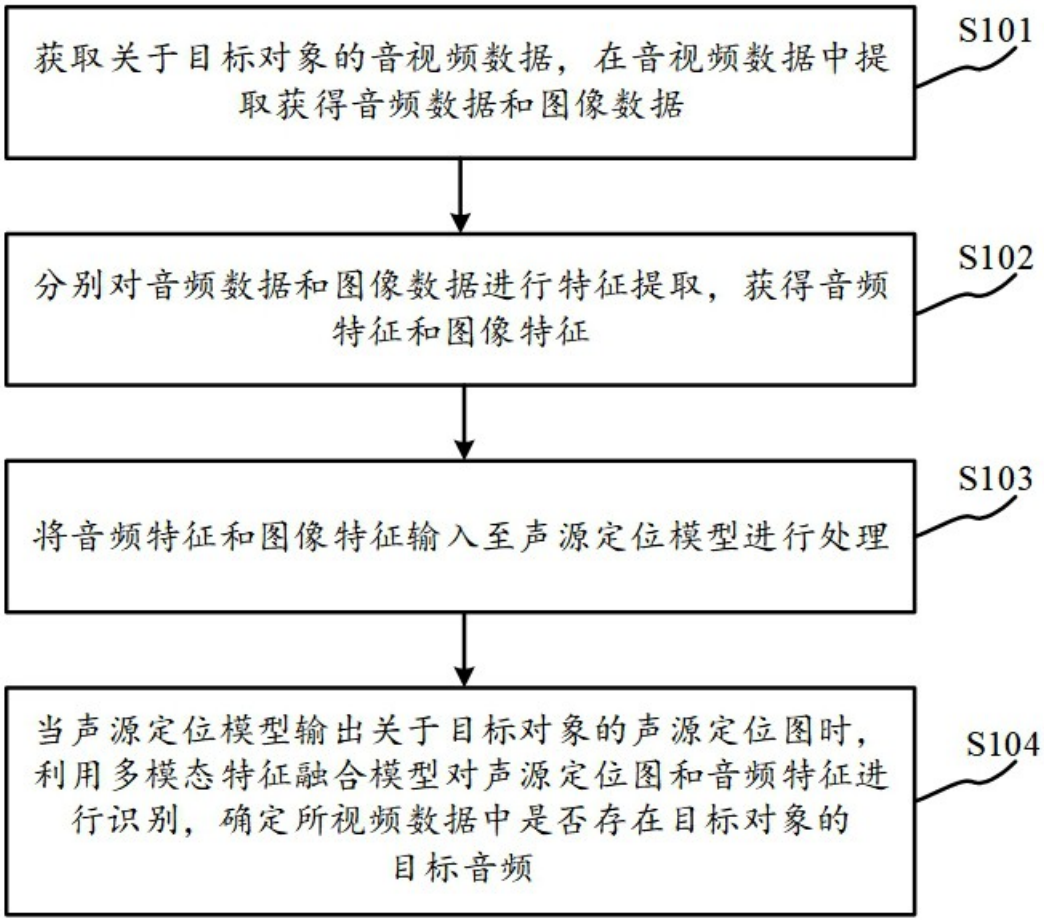

获取关于目标对象的音视频数据，在音视频数据中提取获得音频数据和图像数据
S101
分别对音频数据和图像数据进行特征提取，获得音频特征和图像特征
S102
将音频特征和图像特征输入至声源定位模型进行处理
S103
当声源定位模型输出关于目标对象的声源定位图时，利用多模态特征融合模型对声源定位图和音频特征进行识别，确定所视频数据中是否存在目标对象的目标音频
S104

图1

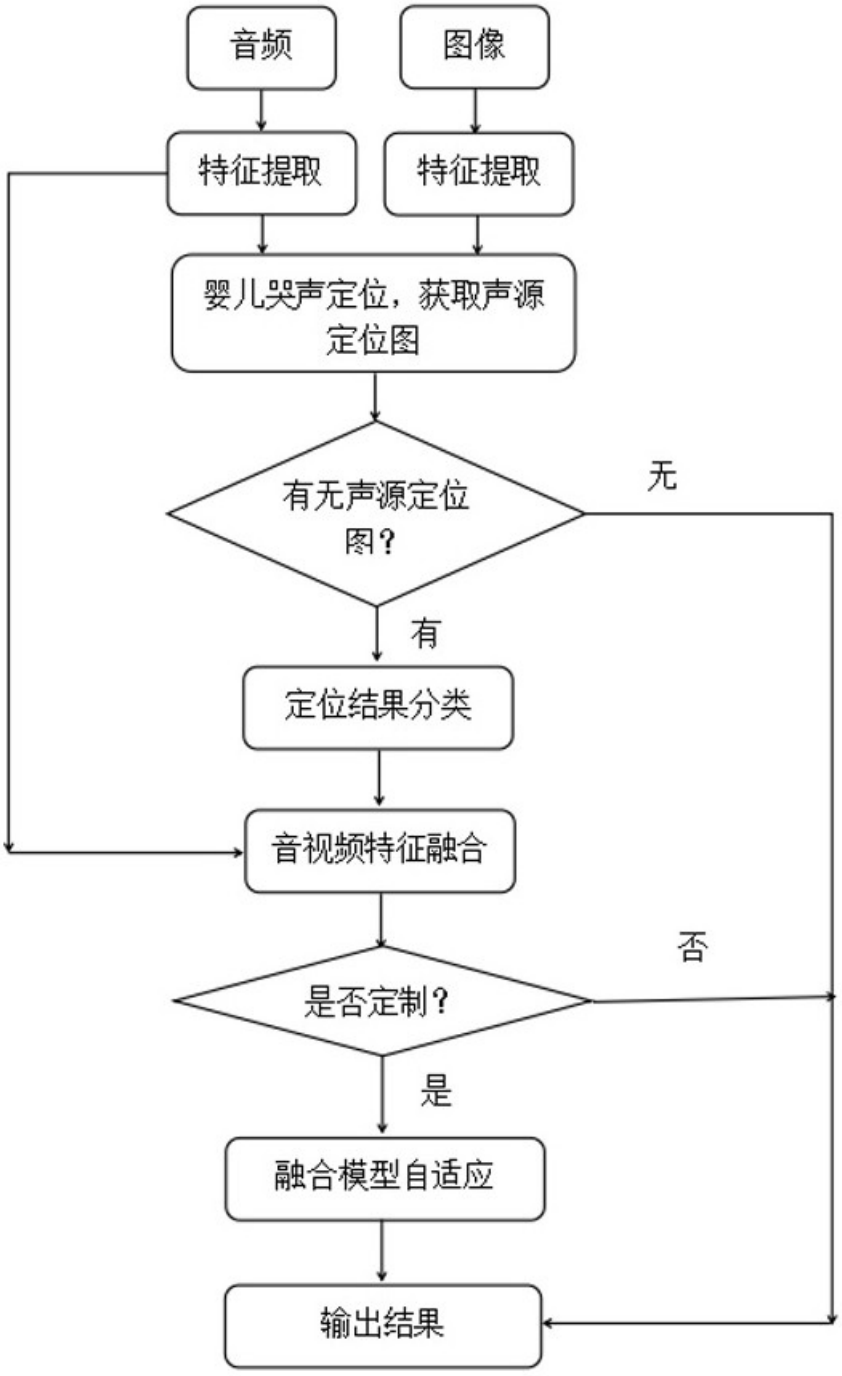

音频
图像
特征提取
特征提取
婴儿哭声定位，获取声源定位图
有无声源定位图？
无
有
定位结果分类
音视频特征融合
是否定制？
否
是
融合模型自适应
输出结果

图2

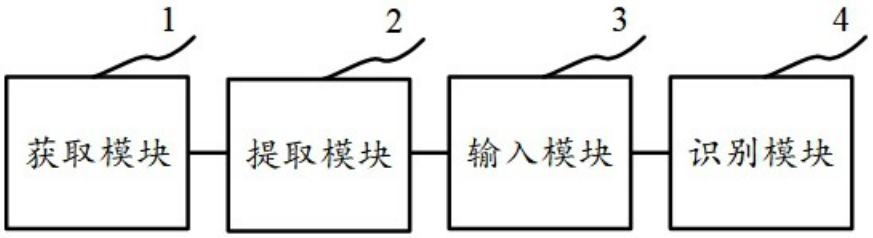


图3

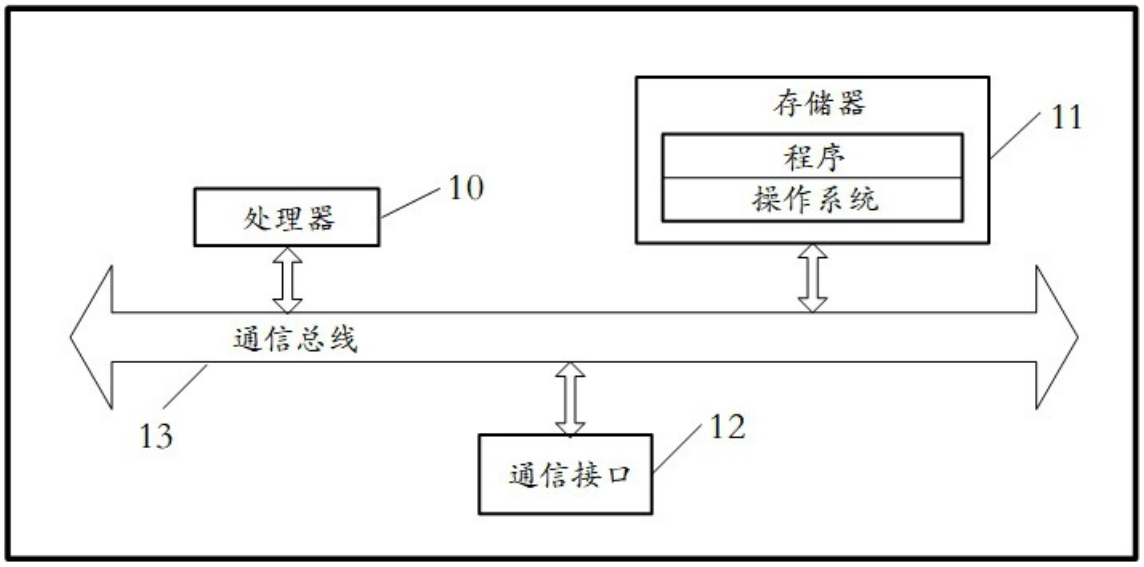


图4